MaxViT: Multi-Axis Vision Transformer

Year: 2022 | Citations: 855 | Authors: Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, P. Milanfar

---

Abstract

Transformers have recently gained significant attention in the computer vision community. However, the lack of scalability of self-attention mechanisms with respect to image size has limited their wide adoption in state-of-the-art vision backbones. In this paper we introduce an efficient and scalable attention model we call multi-axis attention, which consists of two aspects: blocked local and dilated global attention. These design choices allow global-local spatial interactions on arbitrary input resolutions with only linear complexity. We also present a new architectural element by effectively blending our proposed attention model with convolutions, and accordingly propose a simple hierarchical vision backbone, dubbed MaxViT, by simply repeating the basic building block over multiple stages. Notably, MaxViT is able to "see" globally throughout the entire network, even in earlier, high-resolution stages. We demonstrate the effectiveness of our model on a broad spectrum of vision tasks. On image classification, MaxViT achieves state-of-the-art performance under various settings: without extra data, MaxViT attains 86.5% ImageNet-1K top-1 accuracy; with ImageNet-21K pre-training, our model achieves 88.7% top-1 accuracy. For downstream tasks, MaxViT as a backbone delivers favorable performance on object detection as well as visual aesthetic assessment. We also show that our proposed model expresses strong generative modeling capability on ImageNet, demonstrating the superior potential of MaxViT blocks as a universal vision module. The source code and trained models will be available at https://github.com/google-research/maxvit.